PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/00, 1/68, C07H 21/00

A1

(11) International Publication Number: WO 97/27317

(43) International Publication Date: 31 July 1997 (31.07.97)

(21) International Application Number: PCT/US97/01603

(22) International Filing Date: 22 January 1997 (22.01.97)

(30) Priority Data:

60/010,471	23 January 1996 (23.01.96)	US
Not furnished	9 January 1997 (09.01.97)	US

(60) Parent Application or Grant
(63) Related by Continuation
US 60/010,471 (CIP)
Filed on 23 January 1996 (23.01.96)

(71) Applicant *(for all designated States except US)*: AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: LOCKHART, David, J. [US/US]; Apartment 205, 480 Oak Grove Drive, Santa Clara, CA 95054 (US). CHEE, Mark [AU/US]; 3199 Waverly Street, Palo Alto, CA 94306 (US). GUNDERSON, Kevin [US/US]; 1090 Tanland Drive 103, Palo Alto, CA 94303 (US). LAI, Chaoqiang [CN/US]; 1991 Halford Avenue #230, Santa Clara, CA 95051 (US). WODICKA, Lisa [US/US]; 3770 Flora Vista #603, Santa Clara, CA 95051 (US). CRONIN, Maureen, T. [US/US]; 771 Anderson Drive, Los Altos, CA 94024 (US). LEE, Danny [US/US]; 5520 Le Franc Drive, San Jose, CA 95118 (US). TRAN, Huu, M. [US/US]; 3697 Cape Cod Court #1, San Jose, CA 95117 (US). MATSUZAKI, Hajime [US/US]; 562 Kendall Avenue #26, Palo Alto, CA 94306 (US). McGALL, Glenn, H. [CA/US]; 750 North Shoreline Boulevard, Mountain View, CA 94041 (US). BARONE, Anthony, D. [US/US]; 2118 Ellen Avenue, San Jose, CA 95125 (US).

(74) Agents: HUNTER, Tom et al.; Townsend and Townsend and Crew L.L.P., 8th floor, Two Embarcadero Center, San Francisco, CA 94111 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: NUCLEIC ACID ANALYSIS TECHNIQUES

(57) Abstract

The present invention provides a simplified method for identifying differences in nucleic acid abundances (e.g., expression levels) between two or more samples. The methods involve providing an array containing a large number (e.g. greater then 1,000) of arbitrarily selected different oligonucleotide probes where the sequence and location of each different probe is known. Nucleic acid samples (e.g. mRNA) from two or more samples are hybridized to the probe arrays and the pattern of hybridization is detected. Differences in the hybridization patterns between the samples indicates differences in expression of various genes between those samples. This invention also provides a method of end-labeling a nucleic acid. In one embodiment, the method involves providing a nucleic acid, providing a labeled oligonucleotide and then enzymatically ligating the oligonucleotide to the nucleic acid. Thus, for example, where the nucleic acid is an RNA, a labeled oligoribonucleotide can be ligated using an RNA ligase. In another embodiment, the end labeling can be accomplished by providing a nucleic acid, providing labeled nucleoside triphosphates, and attaching the nucleoside triphosphates to the nucleic acid using a terminal transferase.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

NUCLEIC ACID ANALYSIS TECHNIQUES

CROSS REFERENCE TO RELATED APPLICATIONS

This is a continuation-in-part of U.S.S.N. 60/010,471 filed on January 23, 1996 and a continuation-in-part of provisional patent application for "Labeling of Nucleic Acids" naming Lockhart, Cronin, Lee, Tran, Matsuzaki, McGall and Barone as inventors, filed on January 9, 1997, both of which are herein incorporated by reference for all purposes.

BACKGROUND OF THE INVENTION

Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (*e.g.* through control of initiation, provision of RNA precursors, RNA processing, *etc.*) of particular genes. For example, losses and gains of genetic material play an important role in malignant transformation and progression. These gains and losses are thought to be "driven" by at least two kinds of genes. Oncogenes are positive regulators of tumorigenesis, while tumor suppressor genes are negative regulators of tumorigenesis (Marshall, *Cell*, 64: 313-326 (1991); Weinberg, Science, 254: 1138-1146 (1991)). Therefore, one mechanism of activating unregulated growth is to increase the number of genes coding for oncogene proteins or to increase the level of expression of these oncogenes (*e.g.* in response to cellular or environmental changes), and another is to lose genetic material or to decrease the level of expression of genes that code for tumor suppressors. This model is supported by the losses and gains of genetic material associated with glioma progression (Mikkelson *et al. J. Cell. Biochem.* 46: 3-8 (1991)). Thus, changes in the expression (transcription) levels of particular genes

(*e.g.* oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

Similarly, control of the cell cycle and cell development, as well as diseases, are characterized by the variations in the transcription levels of particular genes. Thus, for example, a viral infection is often characterized by the elevated expression of genes of the particular virus. For example, outbreaks of *Herpes simplex*, Epstein-Barr virus infections (*e.g.* infectious mononucleosis), cytomegalovirus, Varicella-zoster virus infections, parvovirus infections, human papillomavirus infections, *etc.* are all characterized by elevated expression of various genes present in the respective virus. Detection of elevated expression levels of characteristic viral genes provides an effective diagnostic of the disease state. In particular, viruses such as herpes simplex, enter quiescent states for periods of time only to erupt in brief periods of rapid replication. Detection of expression levels of characteristic viral genes allows detection of such active proliferative (and presumably infective) states.

The use of "traditional" hybridization protocols for monitoring or quantifying gene expression is problematic. For example two or more gene products of approximately the same molecular weight will prove difficult or impossible to distinguish in a Northern blot because they are not readily separated by electrophoretic methods. Similarly, as hybridization efficiency and cross-reactivity varies with the particular subsequence (region) of a gene being probed it is difficult to obtain an accurate and reliable measure of gene expression with one, or even a few, probes to the target gene.

The development of VLSIPS™ technology provided methods for synthesizing arrays of many different oligonucleotide probes that occupy a very small surface area. *See* U.S. Patent No. 5,143,854 and PCT patent publication No. WO 90/15070. U.S. Patent application Serial No. 082,937, filed June 25, 1993, describes methods for making arrays of oligonucleotide probes that can be used to provide the complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific nucleotide sequence.

Previous methods of measuring nucleic acid abundance differences or changes in the expression of various genes (*e.g.*, differential diaplay, SAGE, cDNA sequencing, clone spotting, *etc.*) require assumptions about, or prior knowledge regarding

the target sequences in order to design appropriate sequence-specific probes. Other methods, such as subtractive hybridization, do not require prior sequence knowledge, but also do not directly provide sequence information regarding differentially expressed nucleic acids.

Summary of the Invention

The present invention, in one embodiment, provides methods of monitoring the expression of a multiplicity of preselected genes (referred to herein as "expression monitoring"). In another embodiment this invention provides a way of identifying differences in the compositions of two or more nucleic acid (*e.g.*, RNA or DNA) samples. Where the nucleic acid abundances reflect expression levels in biological samples from which the samples are derived, the invention provides a method for identifying differences in expression profiles bewteen two or more samples. These "generic difference screening methods" are rapid, simple to apply, require no *a priori* assumptions regarding the particular sequences whose expression may differ between the two samples, and provide direct sequence information regarding the nucleic acids whose abundances differ between the samples.

In one embodiment, this invention provides a method of identifying differences in nucleic acid levels between two or more nucleic acid samples. The method involves the steps of: (a) providing one or more oligonucleotide arrays said arrays comprising probe oligonucleotides attached to a surface; (b) hybridizing said nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid samples and probe oligonucleotides in said one or more arrays that are complementary to said nucleic acids or subsequences thereof;(c) contacting said one or more arrays with a nucleic acid ligase; and (d) determining differences in hybridization between said nucleic acid samples wherein said differences in hybridization indicate differences in said nucleic acid levels.

In another embodiment, the method of identifying differences in nucleic acid levels between two or more nucleic acid samples involves the steps of: (a) providing one or more oligonucleotide arrays comprising probe oligonucleotides wherein said probe oligonucleotides comprise a constant region and a variable region; (b) hybridizing said

nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid samples and said variable regions that are complementary to said nucleic acids or subsequences thereof; and (c) determining differences in hybridization between said nucleic acid samples wherein said differences in hybridization indicate differences in said nucleic acid levels.

In yet another embodiment, the method of identifying differences in nucleic acid levels between two or more nucleic acid samples involves the steps of: (a) providing one or more high density oligonucleotide arrays; (b) hybridizing said nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid samples and probe oligonucleotides in said one or more arrays that are complementary to said nucleic acids or subsequences thereof; and (c) determining the differences in hybridization between said nucleic acid samples wherein said differences in hybridization indicate differences in said nucleic acid levels.

In still yet another embodiment, the method of identifying differences in nucleic acid levels between two or more nucleic acid samples involves the steps of: (a) providing one or more oligonucleotide arrays each comprising probe oligonucleotides wherein said probe oligonucleotides are not chosen to hybridize to nucleic acids derived from particular preselected genes or mRNAs; (b) hybridizing said nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid samples and probe oligonucleotides in said one or more arrays that are complementary to said nucleic acids or subsequences thereof; and (d) determining differences in hybridization between said nucleic acid samples wherein said differences in hybridization indicate differences in said nucleic acid levels.

In another embodiment, the methods of identifying differences in nucleic acid levels between two or more nucleic acid samples involves the steps of: (a) providing one or more oligonucleotide arrays each comprising probe oligonucleotides wherein said probe oligonucleotides comprise a nucleotide sequences or subsequences selected according to a process selected from the group consisting of a random selection, a haphazard selection, a nucleotide composition biased selection, and all possible oligonucleotides of a preselected length; (b) hybridizing said nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid

samples and probe oligonucleotides in said one or more arrays that are complementary to said nucleic acids or subsequences thereof; and (c) determining differences in hybridization between said nucleic acid samples wherein said differences in hybridization indicate differences in said nucleic acid levels.

In another embodiment, the methods of identifying differences in nucleic acid levels between two or more nucleic acid samples involve the steps of: (a) providing one or more oligonucleotide arrays each comprising probe oligonucleotides wherein said probe oligonucleotides comprise a nucleotide sequence or subsequences selected according to a process selected from the group consisting of a random selection, a haphazard selection, a nucleotide composition biased selection, and all possible oligonucleotides of a preselected length; (b) providing software describing the location and sequence of probe oligonucleotides on said array; (c) hybridizing said nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid samples and probe oligonucleotides in said one or more arrays that are complementary to said nucleic acids or subsequences thereof; and (d) operating said software such that said hybridizing indicates differences in said nucleic acid levels.

This invention also provides methods of simultaneously monitoring the expression of a multiplicity of genes. In one embodiment these methods involve (a) providing a pool of target nucleic acids comprising RNA transcripts of one or more of said genes, or nucleic acids derived from said RNA transcripts; (b) hybridizing said pool of nucleic acids to an oligonucleotide array comprising probe oligonucleotides immobilized on a surface; (c) contacting said oligonucleotide array with a ligase; and (d) quantifying the hybridization of said nucleic acids to said array wherein said quantifying provides a measure of the levels of transcription of said genes.

Still yet another method of identifying differences in nucleic acid levels between two or more nucleic acid samples involves the steps of: (a) providing one or more arrays of oligonucleotides each array comprising pairs of probe oligonucleotides where the members of each pair of probe oligonucleotides differ from each other in preselected nucleotides; (b) hybridizing said nucleic acid samples to said one or more arrays to form hybrid duplexes between nucleic acids in said nucleic acid samples and probe oligonucleotides in said one or more arrays that are complementary to said nucleic acids or

subsequences thereof; (c) determining the differences in hybridization between said nucleic acid samples wherein said differences in hybridization indicate differences in said nucleic acid levels.

Another method of simultaneously monitoring the expression of a multiplicity of genes, involves the steps of: (a) providing one or more oligonucleotide arrays comprising probe oligonucleotides wherein said probe oligonucleotides comprise a constant region and a variable region; (b) providing a pool of target nucleic acids comprising RNA transcripts of one or more of said genes, or nucleic acids derived from said RNA transcripts; (c) hybridizing said pool of nucleic acids to an array of oligonucleotide probes immobilized on a surface; and (d) quantifying the hybridization of said nucleic acids to said array wherein said quantifying provides a measure of the levels of transcription of said genes.

This invention additionally provides methods of making a nucleic acid array for identifying differences in nucleic acid levels between two or more nucleic acid samples. In one embodiment the method involves thesteps of: (a) providing an oligonucleotide array comprising probe oligonucleotides wherein said probe oligonucleotides comprise a constant region and a variable region; (b) hybridizing one or more of said nucleic acid samples to said arrays to form hybrid duplexes of said variable region and nucleic acids in said nucleic acid samples comprising subsequences complementary to said variable region; (c) attaching the sample nucleic acids comprising said hybrid duplexes to said array of probe oligonucleotides; and (d) removing unattached nucleic acids to provide a high density oligonucleotide array bearing sample nucleic acids attached to said array.

In another embodiment the method of making a nucleic acid array for identifying differences in nucleic acid levels between two or more nucleic acid samples, involves the steps of: (a) providing a high density array; (b) contacting said array one or more of said two or more nucleic acid samples whereby nucleic acids of said one of said two or more nucleic acid samples form hybrid duplexes with probe oligonucleotides in said arrays; (c) attaching the sample nucleic acids comprising said hybrid duplexes to said array of probe oligonucleotides; and (d) removing unattached nucleic acids to provide a high density oligonucleotide array bearing sample nucleic acids attached to said array.

This invention additionally provides kits for practice of the methods described herein. One kit comprises a container containing one or more oligonucleotide arrays said arrays comprising probe oligonucleotides attached to a surface; and a container containing a ligase. Another kit comprises a container containing one or more oligonucleotide arrays said arrays comprising probe oligonucleotides wherein said probe oligonucleotides comprise a constant region and a variable region. This kit optionally includes a constant oligonucletide complementary to said constant region or a subsequence thereof.

Preferred high density oligonucleotide arrays of this invention comprise more than 100 different probe oligonucleotides wherein: each different probe oligonucleotide is localized in a predetermined region of the array; each different probe oligonucleotide is attached to a surface through a terminal covalent bond; and the density of said probe different oligonucleotides is greater than about 60 different oligonucleotides per 1 cm^2. The high density arrays can be used in all of the array-based methods discussed herein. High density arrays used for expressio monitoring will typically include oligonucleotide probes selected to be complementary to a nucleic acid derived from one or more preselected genes. In contrast, generic difference screening arrays may contain probe oligonucleotides selected randomly, haphazardly, arbitrarily, or including sequences or subsequences comprising all possible nucleic acid sequences of a particular (preselected) length.

In a preferred embodiment, pools of oligonucleotides or oligonucleotide subsequences comprising all possible nucleic acids of a particular length are selected from the group consisting of all possible 6 mers, all possible 7 mers, all possible 8 mers, all possible 9 mers, all possible 10 mers, all possible 11 mers, and all possible 12 mers

This invention also provides methods of labeling a nucleic acid. In one embodiment, this method involves the steps of: (a) providing a nucleic acid; (b) amplifying said nucleic acid to form amplicons; (c) fragmenting said amplicons to form fragments of said amplicons; and (d) coupling a labeled moiety to at least one of said fragments.

In another embodiment, the methods involve the steps of: (a) providing a nucleic acid; (b) transcribing said nucleic acid to formed a transcribed nucleic acid; (c)

fragmenting said transcribed nucleic acid to form fragments of said transcribed nucleic acid; and (d) coupling a labeled moiety to at least one of said fragments.

In yet another embodiment, the methods involve the steps of: (a) providing at least one nucleic acid coupled to a support; (b) providing a labeled moiety capable of being coupled with a terminal transferase to said nucleic acid; (c) providing said terminal transferase; and (d) coupling said labeled moiety to said nucleic acid using said terminal transferase.

In still another embodiment, the methods involve the steps of: (a) providing at least two nucleic acids coupled to a support; (b) increasing the number of monomer units of said nucleic acids to form a common nucleic acid tail on said at least two nucleic acids; (c) providing a labeled moiety capable of recognizing said common nucleic acid tails; and (d) contacting said common nucleic acid tails and said labeled moiety.

In still yet another embodiment, the methods involve the steps of: (a) providing at least one nucleic acid coupled to a support; (b) providing a labeled moiety capable of being coupled with a ligase to said nucleic acid; (c) providing said ligase; and (d) coupling said labeled moiety to said nucleic acid using said ligase.

This invention also provides compounds of the formulas described herein.

Definitions.

An array of oligonucleotides as used herein refers to a multiplicity of different (sequence) oligonucleotides attached (preferably through a single terminal covalent bond) to one or more solid supports where, when there is a multiplicity of supports, each support bears a multiplicity of oligonucleotides. The term "array" can refer to the entire collection of oligonucleotides on the support(s) or to a subset thereof. The term "same array" when used to refer to two or more arrays is used to mean arrays that have substantially the same oligonucleotide species thereon in substantially the same abundances. The spatial distribution of the oligonucleotide species may differ between the two arrays, but, in a preferred embodiment, it is substantially the same. It is recognized that even where two arrays are designed and synthesized to be identical there are variations in the abundance, composition, and distribution of oligonucleotide probes. These

variations are preferably insubstantial and/or compensated for by the use of controls as described herein.

The phrase "massively parallel screening" refers to the simultaneous screening of at least about 100, preferably about 1000, more preferably about 10,000 and most preferably about 1,000,000 different nucleic acid hybridizations.

The terms "nucleic acid" or "nucleic acid molecule" refer to a deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, and unless otherwise limited, would encompass known analogs of natural nucleotides that can function in a similar manner as naturally occurring nucleotides.

An oligonucleotide is a single-stranded nucleic acid ranging in length from 2 to about 1000 nucleotides, more typically from 2 to about 500 nucleotides in length.

As used herein a "probe" is defined as an oligonucleotide capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, an oligonucleotide probe may include natural (*i.e.* A, G, C, or T) or modified bases (7-deazaguanosine, inosine, *etc.*). In addition, the bases in oligonucleotide probe may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, oligonucleotide probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages.

The term "target nucleic acid" refers to a nucleic acid (often derived from a biological sample and hence referred to also as a sample nucleic acid), to which the oligonucleotide probe specifically hybridizes. It is recognized that the target nucleic acids can be derived from essentially any source of nucleic acids (*e.g.*, including, but not limited to chemical syntheses, amplification reactions, forensic samples, *etc.*) It is either the presence or absence of one or more target nucleic acids that is to be detected, or the amount of one or more target nucleic acids that is to be quantified. The target nucleic acid(s) that are detected preferentially have nucleotide sequences that are complementary to the nucleic acid sequences of the corresponding probe(s) to which they specifically bind (hybridize). The term target nucleic acid may refer to the specific subsequence of a larger nucleic acid to which the probe specifically hybridizes, or to the overall sequence (*e.g.*,

gene or mRNA) whose abundance (concentration) and/or expression level it is desired to detect. The difference in usage will be apparent from context.

A "ligatable oligonucleotide" or "ligatable probe" or "ligatable oligonucleotide probe" refers to an oligonucleotide that is capable of being ligated to another oligonucleotide by the use of a ligase (*e.g.*, T4 DNA ligase). The ligatable oligonucleotide is preferably a deoxyribonucleotide. The nucleotides comprising the ligatable oligonucleotide are preferably the "standard" nucleotides; A, G, C, and Tor U. However derivatized, modified, or alternative nucleotides (*e.g.*, inosine) can be present as long as their presence does not interfere with the ligation reaction. The ligatable probe may be labeled or otherwise modified as long as the label does not interfere with the ligation reaction. Similarly the internucleotide linkages can be modified as long as the modification does not interfere with ligation. Thus, in some instances, the ligatable oligonucleotide can be a peptide nucleic acid.

"Subsequence" refers to a sequence of nucleic acids that comprises a part of a longer sequence of nucleic acids.

A "wobble" refers to a degeneracy at a particular position in an oligonucleotide. A fully degenerate or "4 way" wobble refers to a collection of nucleic acids (*e.g.*, oligonucleotide probes having A, G, C, or T for DNA or A, G, C, or U for RNA at the wobble position.) A wobble may be approximated by the replacement of the nucleotide with inosine which will base pair with A, G, C, or T or U. Typically oligonucleotides containing a fully degenerate wobble produced during chemical synthesis of an oligonucleotide is prepared by using a mixture of four different nucleotide monomers at the particular coupling step in which the wobble is to be introduced.

The term"cross-linking" when used in reference to cross-linking nucleic acids refers to attaching nucleic acids such that they are not separated under typical conditions that are used to denature complementary nucleic acid sequences. Crosslinking preferably involves the formation of covalent linkages between the nucleic acids. Methods of cross-linking nucleic acids are described herein.

The phrase "coupled to a support" means bound directly or indirectly thereto including attachment by covalent binding, hydrogen bonding, ionic interaction, hydrophobic interaction, or otherwise.

"Amplicons" are the products of the amplification of nucleic acids by PCR or otherwise.

"Transcribing a nucleic acid" means the formation of a ribonucleic acid from a deoxyribonucleic acid and the converse (the formation of a deoxyribonucleic acid from a ribonucleic acid). A nucleic acid can be transcribed by DNA-dependent RNA polymerase, reverse transcriptase, or otherwise.

A labeled moiety means a moiety capable of being detected by the various methods discussed herein or known in the art.

The term "complexity"is used here according to standard meaning of this term as established by Britten *et al. Methods of Enzymol.* 29:363 (1974). See, also *Cantor and Schimmel Biophysical Chemistry: Part III* at 1228-1230 for further explanation of nucleic acid complexity.

"Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

The phrase "hybridizing specifically to", refers to the binding, duplexing, or hybridizing of a molecule preferentially to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. The term "stringent conditions" refers to conditions under which a probe will hybridize preferrentially to its target subsequence, and to a lesser extent to, or not at all to, other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength, pH, and nucleic acid concentration) at which 50% of the probes complementary to the target sequence hybridize to the target sequence at equilibrium. (As the target sequences are generally present in excess, at T_m, 50% of the probes are occupied at equilibrium). Typically, stringent conditions will be those in which the salt concentration is at least about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50

nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The perfect match (PM) probe can be a "test probe", a "normalization control" probe, an expression level control probe and the like. A perfect match control or perfect match probe is, however, distinguished from a "mismatch control" or "mismatch probe." In the case of expression monitoring arrays, perfect match probes are typically preselected (designed) to be complementary to particular sequences or subsequences of target nucleic acids (*e.g.*, particular genes). In contrast, in generic difference screening arrays, the particular target sequences are typically unknown. In the latter case, prefect match probes cannot be preselected. The term perfect match probe in this context is to distinguish that probe from a corresponding "mismatch control" that differs from the perfect match in one or more particular preselected nucleotides as described below.

The term "mismatch control" or "mismatch probe", in expression monitoring arrays, refers to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in a high-density array there preferably exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence. In "generic" (*e.g.*, random, arbitrary, haphazard, *etc.*) arrays, since the target nucleic acid(s) are unknown perfect match and mismatch probes cannot be *a priori* determined, designed, or selected. In this instance, the probes are preferably provided as pairs where each pair of probes differ in one or more preselected nucleotides. Thus, while it is not known *a priori* which of the probes in the pair is the perfect match, it is known that when one probe specifically hybridizes to a particular target sequence, the other probe of the pair will act as a mismatch control for that target sequence. It will be appreciated that the perfect match and mismatch probes need not be provided as pairs, but may be provided as larger collections (*e.g.*, 3. 4, 5, or more) of probes that differ from each other in particular preselected nucleotides. While the mismatch(s) may be located anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization

of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions. In a particularly preferred embodiment, perfect matches differ from mismatch controls in a single centrally-located nucleotide.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding, or other interactions, between the labeled target nucleic acids and components of the oligonucleotide array (*e.g.*, the oligonucleotide probes, control probes, the array substrate, *etc.*). Background signals may also be produced by intrinsic fluorescence of the array components themselves. A single background signal can be calculated for the entire array, or a different background signal may be calculated for each region of the array. In a preferred embodiment, background is calculated as the average hybridization signal intensity for the lowest 1% to 10% of the probes in the array, or region of the array. In expression monitoring arrays (*i.e.*, where probes are preselected to hybridize to specific nucleic acids (genes)), a different background signal may be calculated for each target nucleic acid. Where a different background signal is calculated for each target gene, the background signal is calculated for the lowest 1% to 10% of the probes for each gene. Of course, one of skill in the art will appreciate that where the probes to a particular gene hybridize well and thus appear to be specifically binding to a target sequence, they should not be used in a background signal calculation. Alternatively, background may be calculated as the average hybridization signal intensity produced by hybridization to probes that are not complementary to any sequence found in the sample (*e.g.* probes directed to nucleic acids of the opposite sense or to genes not found in the sample such as bacterial genes where the sample is of mammalian origin). Background can also be calculated as the average signal intensity produced by regions of the array that lack any probes at all.

The term "quantifying" when used in the context of quantifying nucleic acid abundances or concentrations (*e.g.*, transcription levels of a gene) can refer to absolute or to relative quantification. Absolute quantification may be accomplished by inclusion of known concentration(s) of one or more target nucleic acids (*e.g.* control nucleic acids such as *BioB* or with known amounts the target nucleic acids themselves) and referencing the

hybridization intensity of unknowns with the known target nucleic acids (*e.g.* through generation of a standard curve). Alternatively, relative quantification can be accomplished by comparison of hybridization signals between two or more genes, or between two or more treatments to quantify the changes in hybridization intensity and, by implication, transcription level.

The"percentage of sequence identity" or "sequence identity" is determined by comparing two optimally aligned sequences or subsequences over a comparison window or span, wherein the portion of the polynucleotide sequence in the comparison window may optionally comprise additions or deletions (*i.e.*, gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical subunit (*e.g.* nucleic acid base or amino acid residue) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Percentage sequence identity when calculated using the programs GAP or BESTFIT (see below) is calculated using default gap weights.

Methods of alignment of sequences for comparison are well known in the art. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman, *Adv. Appl. Math.* 2: 482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48: 443 (1970), by the search for similarity method of Pearson and Lipman, *Proc. Natl. Acad. Sci. USA* 85: 2444 (1988), by computerized implementations of these algorithms (including, but not limited to CLUSTAL in the PC/Gene program by Intelligenetics, Moutain View, California, GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, Wisconsin, USA), or by inspection. In particular, methods for aligning sequences using the CLUSTAL program are well described by Higgins and Sharp in *Gene*, 73: 237-244 (1988) and in *CABIOS* 5: 151-153 (1989)).

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows a schematic of expression monitoring using oligonucleotide arrays. Extracted poly $(A)^+$ RNA is converted to cDNA, which is then transcribed in the presence of labeled ribonucleotide triphosphates. L is either biotin or a dye such as fluorescein. RNA is fragmented with heat in the presence of magnesium ions. Hybridizations are carried out in a flow cell that contains the two-dimensional DNA probe arrays. Following a brief washing step to remove unhybridized RNA, the arrays are scanned using a scanning confocal microscope. Alternatives in which cellular mRNA is directly labeled without a cDNA intermediate are described in the Examples. Image analysis software converts the scanned array images into text files in which the observed intensities at specific physical locations are associated with particular probe sequences.

Fig. 2A shows a fluorescent image of a high density array containing over 16,000 different oligonucleotide probes. The image was obtained following hybridization (15 hours at 40°C) of biotin-labeled randomly fragmented sense RNA transcribed from the murine B cell (Tl0) cDNA library, and spiked at the level of 1:3,000 (50 pM equivalent to about 100 copies per cell) with 13 specific RNA targets. The brightness at any location is indicative of the amount of labeled RNA hybridized to the particular oligonucleotide probe. Fig. 2B shows a small portion of the array (the boxed region of Fig. 2A) containing probes for IL-2 and IL-3 RNAs. For comparison, Fig. 2C shows shown the same region of the array following hybridization with an unspiked T10 RNA samples (T10 cells do not express IL-2 and IL-3). The variation in the signal intensity was highly reproducible and reflected the sequence dependence of the hybridization efficiencies. The central cross and the four corners of the array contain a control sequence that is complementary to a biotin-labeled oligonucleotide that was added to the hybridization solution at a constant concentration (50 pM). The sharpness of the images near the boundaries of the features was limited by the resolution of the reading device (11.25 μm) and not by the spatial resolution of the array synthesis. The pixels in the border regions of each synthesis feature were systematically ignored in the quantitative analysis of the images.

Fig. 3 provides a log/log plot of the hybridization intensity (average of the PM-MM intensity differences for each gene) versus concentration for 11 different RNA targets. The hybridization signals were quantitatively related to target concentration. The

experiments were performed as described in the Examples herein and in Fig. 2. The ten 10 cytokine RNAs (plus *bioB*) were spiked into labeled T10 RNA at levels ranging from 1:300,000 to 1:3,000. The signals continued to increase with increased concentration up to frequencies of 1:300, but the response became sublinear at the high levels due to saturation of the probe sites, The linear range can be extended to higher concentrations by using shorter hybridization times. RNAs from genes expressed in T10 cells (IL-10, β-actin and GAPDH) were also detected at levels consistent with results obtained by probing cDNA libraries.

Fig. 4 shows cytokine mRNA levels in the murine 2D6 T helper cell line at different times following stimulation with PMA and a calcium ionophore. Poly $(A)^+$ RNA was extracted at 0, 2, 6, and 24 hours following stimulation and converted to double stranded cDNA containing an RNA polymerase promoter. The cDNA pool was then transcribed in the presence of biotin labeled ribonucleotide triphosphates, fragmented, and hybridized to the oligonucleotide probe arrays for 2 and 22 hours. The fluorescence intensities were converted to RNA frequencies by comparison with the signals obtained for a bacterial RNA (biotin synthetase) spiked into the samples at known amounts prior to hybridization. A signal of 50,000 corresponds to a frequency of approximately 1:100,000 to a frequency of 1:5,000, and a signal of 100 to a frequency of 1:50,000. RNAs for IL-2, IL-4, IL-6, and IL-12p40 were not detected above the level of approximately 1:200,000 in these experiments. The error bars reflect the estimated uncertainty (25 percent) in the level for a given RNA relative to the level for the same RNA at a different time point. The relative uncertainty estimate was based on the results of repeated spiking experiments, and on repeated measurements of IL-10, β-actin and GAPDH RNAs in preparations from both T10 and 2D6 cells (unstimulated). The uncertainty in the absolute frequencies includes message-to-message differences in the hybridization efficiency as well as differences in the mRNA isolation, cDNA synthesis, and RNA synthesis and labeling steps. The uncertainty in the absolute frequencies is estimated to be a factor of three.

Fig. 5 shows a fluorescence image of an array containing over 63,000 different oligonucleotide probes for 118 genes. The image was obtained following overnight hybridization of a labeled murine B cell RNA sample. Each square synthesis region is 50 x 50 μm and contains 107 to 108 copies of a specific oligonucleotide. The

array was scanned at a resolution of 7.5 μm in approximately 15 minutes. The bright rows indicate RNAs present at high levels. Lower level RNAs were unambiguously detected based on quantitative evaluation of the hybridization patterns. A total of 21 murine RNAs were detected at levels ranging from approximately 1:300,000 to 1:100. The cross in the center, the checkerboard in the corners, and the MUR-1 region at the top contain probes complementary to a labeled control oligonucleotide that was added to all samples.

Fig. 6 shows an example of a computer system used to execute the software of an embodiment of the present invention.

Fig. 7 shows a system block diagram of a typical computer system used to execute the software of an embodiment of the present invention.

Fig. 8 shows the high level flow of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes.

Fig. 9 shows the flow of a process of determining if a gene is expressed utilizing a decision matrix.

Figs. 10A and 10B show the flow of a process of determining the expression of a gene by comparing baseline scan data and experimental scan data.

Fig. 11 shows the flow of a process of increasing the number of probes for monitoring the expression of genes after the number of probes has been reduced or pruned.

Figs. 12a and 12b illustrate the probe oligonucleotide/ligation reaction system. Fig. 12 generally illustrates the various components of the probe oligonucleotide/ligation reaction system. Fig. 12b illustrates discrimination of non-perfectly complementary target:oligonucleotide hybrids using the probe oligonucleotide/ligation reaction system.

Figs. 13a, 13b, 13c, and 13d illustrate the various components of ligation/hybridization reactions and illustrates various ligation strategies. Fig. 13a illustrates various components of the ligation/hybridization reaction some of which are optional in various embodiments. Fig. 13b illustrates a ligatiion strateby that discriminates mismatches at the terminus of the probe oligonucleotide. Fig. 13c illustrates a ligation strategy that discriminates mismatches at the terminus of the sample oligonucleotide. Fig.

13d illustrates a method for improving the discrimination at both the probe terminus and the sample terminus.

Figs.14a, 14b, 14c, and 14d illustrates a ligation discrimination used in conjunction with a restriction digest of the sample nucleic acid. Fig. 14a shows the recognition site and cleavage pattern of SacI (a 6 cutter) and Hsp92 II (4 cutter). Fig. 14b illustrates the effect of SacI cleavage on a (target) nucleic acid sample. Fig. 14c illustrates a 6 Mb genome (*i.e., E. coli*) digested with SacI and SphI generating ~ 1kb genomic fragments with a 5' C. Fig. 14d illustrates the hybridization/ligation of these fragments to a generic difference screening chip and their subsequent use as probes to hybridize to the appropriate ncuelic acid (Format I) or the fragments are labeled, hybridized/ligated to the oligonucletide aray and directly analyzed (Format II).

Figs. 15a, 15b, 15c, 15d, and 15e illustrate the analysis of differntial diaplay DNA fragments on a generic difference screenign array. Fig. 15a shows first strand cDNA synthesis by reverse transcripton of poly(a) mRNA using an anchored poly(T) primer. Fig. 15b illustrates upstream primers for PCR reaction containing an engineered restrictionsite and degenerate bases (N=A,G,C,T) at the 3' end. Fig. 15c shows randomly primed PCR of first strand cDNA. Fig. 15d shows restrictiondigest of PCR products, and Fig. 15e shows sorting of PCR products on a generic gligationarray by their 5'end.

Figs. 16a, 16b, and 16c illustrate the differences between replicate 1 and replicate 2 for sample 1 and sample 2 nucleic acids. Fig. 16a shows the differences between replicate 1 and replicate 2 for sample 1, the normal cell line. Fig. 16b shows the differences between replicate 1 and replicate 2 for sample 2, the tumor cell line). Figure 16c plots the differences between sample 1 and 2 averaged over the two replicates.

Figs. 17a, 17b, and 17c illustrates the data of Figs 16A, 16b, and 16c filtered. Figure 17a shows the relative change in hybridization intensities of replicate 1 and 2 of sample 1 for the difference of each oligonucleotide pair. Fig. 17b shows the ratio of replicate 1 and 2 of sample 2 for the difference of each oligoncleotide pair, normalized, filtered, and plotted the same way as in Figure 17A. Fig. 17c shows the ratio of sample 1 and sample 2 averaged over two replicates for the difference of each oligonucleotide pair. The ratio is calculated as in Fig. 17A, but based on the absolute value of

$[(X_{21k1}+X_{22k2})/2]/[(X_{11k1}+X_{12k2})/2]$ and $[(X_{11k1}+X_{12k2})/2]/[(X_{21k1}+X_{22k2})/2]$ after normalization as in Fig. 16c.

Fig. 18 illustrates post-fragmentation labeling using a CIAP treatment.

Fig. 19 provides a schematic illustration of pos-hybridization end labeling on a high density oligonucleotide array.

Fig. 20 provides a schematic illustration end-labeling utilizing pre-reaction of a high density array prior to hybridization and end labeling.

Fig. 21 illustrates the results of a measure of post-hybridization TdTase end labeling call accuracy.

Fig. 22 illustrates oligo dT labeling on a high density oligonucleotide array.

Fig. 23 illustrates various labeling reagents suitable for use in the methods disclosed herein. Fig. 23a shows various labeling reagents. Fig. 23b shows still other labeling reagents. Fig. 23c shows non-ribose or non-2'-deoxyribose-containing labels. Fig. 23d shows sugar-modified nucleotide analogue labels 23d.

Fig. 24. illustrates resequencing of a target DNA molecule with a set of generic n-mer tiling probes.

Fig. 25 illustrates four tiling arrrays present on a 4-mer generic array.

Fig. 26 illustrates base calling at the 8th position in the target.

Fig. 27 illustrates a base vote table.

Fig. 28 illustrates the effect of applying correctness score transform to HIV data.

Fig. 29 illustrates mutation detection by intensity comparisons.

Fig. 30 illustrates bubble formation detection of mutation in the HIV genome.

Fig. 31 illustrates induced difference nearest neighbor probe scoring.

Fig. 32 illustrates mutations found in an HIV PCR target (B) using a generic ligation GeneChip™ and induced difference analysis.

Fig. 33 illustrates mutation detection using comparisons between a reference target and a sample target.

DETAILED DESCRIPTION

I. Expression Monitoring and Generic Difference Screening.

This invention provides methods of expression monitoring and generic difference screening. The term expression monitoring is used to refer to the determination of levels of expression of particular, typically preselected, genes. In a preferred embodiment, the expression monitoring methods of this invention utilize high density arrays of oligonucleotides selected to be complementary to predetermined subsequences of the gene or genes whose expression levels are to be detected. Nucleic acid samples are hybridized to the arrays and the resulting hybridization signal provides an indication of the level of expression of each gene of interest. Because of the high degree of probe redundancy (typically there are multiple probes per gene) the expression monitoring methods provide an essentially accurate absolute measurement and do not require comparison to a reference nucleic acid.

In another embodiment, this invention provides generic difference screening methods, that identify differences in the abundance (concentration) of particular nucleic acids in two or more nucleic acid samples. The generic difference screening methods involve hybridizing two or more nucleic acid samples to the same array high density oligonucleotide array, or to different high density oligonucleotide arrays having the same oligonucleotide probe composition, and optionally the same oligonucleotide spatial distribution. The resulting hybridizations are then compared allowing determination which nucleic acids differ in abundance (concentration) between the two or more samples.

Where the concentrations of the nucleic acids comprising the samples reflects transcription levels genes in a sample from which the nucleic acids are derived, the generic difference screening methods permit identification of differences in transcription (and by implication in expression) of the nucleic acids comprising the two or more samples. The differentially (*e.g.*, over- or under) expressed nucleic acids thus identified can be used (*e.g.*, as probes) to determine and/or isolate those genes whose expression levels differs between the two or more samples.

The generic difference screening methods are advantageous in that, in contrast to the expression monitoring methods, they require no *a priori* assumptions about the probe oligonucleotide composition of the array. To the contrary, the sequences of the

probe oligonucleotides may be random, haphazard, or any arbitrary subset of oligonucleotide probes. Where the oligonucleotide probes are short enough (*e.g.*, less than or equal to a 12 mer) the array may contain every possible nucleic acid of that length. Despite the fact that the generic difference screening arrays might be arbitrary or random, since the sequence of each probe in the array is known the generic difference screening methods still provide direct sequence information regarding the differentially expressed nucleic acids in the samples.

The expression monitoring and generic difference screening methods of this invention involve providing an array containing a large number (*e.g.* greater than 1,000) of arbitrarily selected different oligonucleotide probes (probe oligonucleotides) where the sequence and location in the array of each different probe is known. Nucleic acid samples (*e.g.* mRNA) are hybridized to the probe arrays and the pattern of hybridization is detected.

It is demonstrated herein and in copending applications U. S Patent Serial No. 08/529,115 filed on September 15, 1995 and PCT/US96/14839 that hybridization with high density oligonucleotide probe arrays provides an effective means of detecting and/or quantifying the expression of particular nucleic acids in complex nucleic acid populations. The expression monitoring and difference screening methods of this invention may be used in a wide variety of circumstances including detection of disease, identification of differential gene expression between two samples (*e.g.*, a pathological as compared to a healthy sample), screening for compositions that upregulate or downregulate the expression of particular genes, and so forth.

In one preferred embodiment, the methods of this invention are used to monitor the expression (transcription) levels of nucleic acids whose expression is altered in a disease state. For example, a cancer may be characterized by the overexpression of a particular marker such as the HER2 (c-erbB-2/neu) proto-oncogene in the case of breast cancer. Similarly, overexpression of receptor tyrosine kinases (RTKs) is associated with the etiology of a number of tumors including carcinomas of the breast, liver, bladder, pancreas, as well as glioblastomas, sarcomas and squamous carcinomas (*see* Carpenter, *Ann. Rev. Biochem.*, 56: 881-914 (1987)). Conversely, a cancer (*e.g.*, colerectal, lung and breast) may be characterized by the mutation of or underexpression of a tumor suppressor

gene such as P53 (*see, e.g.*, Tominaga *et al. Critical Rev. in Oncogenesis*, 3: 257-282 (1992)).

Where the particular genes of interest are known, the high density arrays will preferably contain probe oligonucleotides selected to be complementary to the sequences or subsequences of those genes of interest. High probe redundancy for each gene of interest can be achieved and absolute expression levels of each gene can be determined.

Conversely, where it is unknown which genes differ in expression between the healthy and disease state the generic difference screening methods of this invention are particularly appropriate. Hybridization of the healthy and pathological nucleic acids to the generic difference screening arrays disclosed herein and comparison of the hybridization patterns identifies those genes whose regulation is altered in the pathological state.

Similarly, the expression monitoring and generic difference screening methods of this invention can be used to monitor expression of various genes in response to defined stimuli, such as a drug, cell activation, *etc.* The methods are particularly advantageous because they permit simultaneous monitoring of the expression of large numbers of genes. This is especially useful in drug research if the end point description is a complex one, not simply asking if one particular gene is overexpressed or underexpressed. Thus, where a disease state or the mode of action of a drug is not well characterized, the methods of this invention allow rapid determination of the particularly relevant genes. Again, where the gene of interest is known or suspected, expression monitoring methods will preferably be used, while generic screening methods will be used when the particular genes of interest are unknown.

Using the generic difference screening methods disclosed herein, lack of knowledge regarding the particular genes does not prevent identification of useful therapeutics. For example, if the hybridization pattern on a particular high density array for a healthy cell is known and significantly different from the pattern for a diseased cell, then libraries of compounds can be screened for those that cause the pattern for a diseased cell to become like that for the healthy cell. This provides a very detailed measure of the cellular response to a drug.

Generic difference screening methods thus provide a powerful tool for gene discovery and for elucidating mechanisms underlying complex cellular responses to various stimuli. For example, in one embodiment, generic difference screening can be used for "expression fingerprinting". Suppose it is found that the mRNA from a certain cell type displays a distinct overall hybridization pattern that is different under different conditions (*e.g.* when harboring mutations in particular genes, in a disease state). Then this pattern of expression (an expression fingerprint), if reproducible and clearly differentiable in the different cases can be used as a very detailed diagnostic. It is not even required that the pattern be fully interpretable, but just that it is specific for a particular cell state (and preferably of diagnostic and/or prognostic relevance).

Both expression monitoring methods and generic difference screening may also be used in drug safety studies. For example, if one is making a new antibiotic, then it should not significantly affect the expression profile for mammalian cells. The hybridization pattern could be used as a detailed measure of the effect of a drug on cells. In other words, as a toxicological screen.

The expression monitoring and generic difference screening methods of this invention are particularly well suited for gene discovery. For example, as explained above, the generic difference screening methods identify differences in abundances of nucleic acids in two or more samples. These differences may indicate changes in the expression levels of previously unknown genes. The sequence information provided by a difference screening array can be utilized, as described herein, to identify the unknown gene.

The expression monitoring methods can be used in gene discovery by exploiting the fact that many genes that have been discovered to date have been classified into families based on commonality of the sequences. Because of the extremely large number of probes it is possible to place in the high density array, it is possible to include oligonucleotide probes representing known or parts of known members from every gene class. In utilizing such a "chip" (high density array) genes that are already known would give a positive signal at loci containing both variable and common regions. For unknown genes, only the common regions of the gene family would give a positive signal. The result would indicate the possibility of a newly discovered gene.

The expression monitoring and generic difference screening methods of this invention thus also allow the development of "dynamic" gene databases. The Human Genome Project and commercial sequencing projects have generated large static databases which list thousands of sequences without regard to function or genetic interaction. Analyses using the methods of this invention produces "dynamic" databases that define a gene's function and its interactions with other genes. Without the ability to monitor the expression of large numbers of genes simultaneously, or the abilito to detect differences in abundances of large numbers of "unknown" nucleic acids simultaneously, the work of creating such a database is enormous.

The tedious nature of using DNA sequence analysis for determining an expression pattern involves preparing a cDNA library from the RNA isolated from the cells of interest and then sequencing the library. As the DNA is sequenced, the operator lists the sequences that are obtained and counts them. Thousands of sequences would have to be determined and then the frequency of those gene sequences would define the expression pattern of genes for the cells being studied.

By contrast, using an expression monitoring, or generic difference screening, array to obtain the data according to the methods of this invention is relatively fast and easy. For example to in one embodiment, cells may be stimulated to induce expression. The RNA is obtained from the cells and then either labeled directly or a cDNA copy is created. Fluorescent molecules may be incorporated during the DNA polymerization. Either the labeled RNA or the labeled cDNA is then hybridized to a high density array in one overnight experiment. The hybridization provides a quantitative assessment of the levels of every single one of the hybridized nucleic acids with no additional sequencing. In addition the methods of this invention are much more sensitive allowing a few copies of expressed genes per cell to be detected. This procedure is demonstrated in the examples provided herein. These uses of the methods of this invention are intended to be illustrative and in no manner limiting.

II. High Density Arrays For Generic Difference Screening and Expression Monitoring.

As indicated above, this invention provides methods of monitoring (detecting and/or quantifying) the expression levels of a large number of nucleic acids and/or determining differences in nucleic acid concentrations (abundances) between two or more samples. The methods involve hybridization of one or more a nucleic acid samples (target nucleic acids) to one or more high density arrays of nucleic acid probes and then quantifying the amount of target nucleic acids hybridized to each probe in the array.

While nucleic acid hybridization has been used for some time to determine the expression levels of various genes (*e.g.*, Northern Blot), it was a surprising discovery of this invention that high density arrays are suitable for the quantification of the small variations in abundance (*e.g.*, transcription and, by implication, expression) of a nucleic acid (*e.g.*, gene) in the presence of a large population of heterogenous nucleic acids. The signal (*e.g.*, particular gene or gene product, or differentially abundant nucleic acid) may be present at a concentration of less than about 1 in 1,000, and is often present at a concentration less than 1 in 10,000 more preferably less than about 1 in 50,000 and most preferably less than about 1 in 100,000, 1 in 300,000, or even 1 in 1,000,000.

The oligonucleotide arrays can have oligonucleotides as short as 10 nucleotides, more preferably 15 oligonucleotides and most preferably 20 or 25 oligonucleotides are used to specifically detect and quantify nucleic acid expression levels. Where ligation discrimination methods are used, the oligonculeotide arrays can contain shorter oligonucleotides. In this instance, oligonucleotide arrays comprising oligonucleotides ranging in length from 6 to 15 nucleotides, more preferably from about 8 to about 12 nucleotides in length are preferred. Of course arrays containing longer oligonucleotides, as described herein, are also suitable.

The expression monitoring arrays, which are designed to detect particular preselected genes, provide for simultaneous monitoring of at least about 10, preferably at least about 100, more preferably at least about 1000, still more preferably at least about 10,000, and most preferably at least about 100,000 different genes.

A) Advantages of Oligonucleotide Arrays.

In one preferred embodiment, the high density arrays used in the methods of this invention comprise chemically synthesized oligonucleotides. The use of chemically

synthesized oligonucleotide arrays, as opposed to, for example, blotted arrays of genomic clones, restriction fragments, oligonucleotides, and the like, offers numerous advantages. These advantages generally fall into four categories:

1) Efficiency of production;

2) Reduced intra- and inter-array variability;

3) Increased information content; and

4) Improved signal to noise ratio.

1) Efficiency of production.

In a preferred embodiment, the arrays are synthesized using methods of spatially addressed parallel synthesis (*see, e.g.,* Section V, below). The oligonucleotides are synthesized chemically in a highly parallel fashion covalently attached to the array surface. This allows extremely efficient array production. For example, arrays containing any collection of tens (or even hundreds) of thousands of specifically selected 20 mer oligonucleotides are synthesized in fewer than 80 synthesis cycles. The arrays are designed and synthesized based on sequence information alone. Thus, unlike blotting methods, the array preparation requires no handling of biological materials. There is no need for cloning steps, nucleic acid purifications or amplifications, cataloging of clones or amplification products, and the like. The preferred chemical synthesis of high density oligonucleotide arrays in this invention is thus more efficient than blotting methods and permits the production of highly reproducible high-density arrays.

2) Reduced intra- and inter-array variability.

The use of chemically synthesized high-density oligonucleotide arrays in the methods of this invention improves intra- and inter-array variability. The oligonucleotide arrays preferred for this invention are made in large batches (presently 49 arrays per wafer with multiple wafers synthesized in parallel) in a highly controlled reproducible manner. This makes them suitable as general diagnostic and research tools permitting direct comparisons of assays performed at tifferent times and locations.

Because of the precise control obtainable during the chemical synthesis the arrays of this invention show less than about 25%, preferably less than about 20%, more